15.0 What are the Procedures for Evaluating Model Performance and What is the Role of Diagnostic Analyses?

The results of a model performance evaluation should be considered prior to using modeling to support an attainment demonstration. The performance of an air quality model can be evaluated in two ways: (1) how well is the model able to replicate observed concentrations of ozone and/or precursors (surface and aloft), and (2) how accurate is the model in characterizing sensitivity of ozone to changes in emissions? The first type of evaluation can be broadly classified as an "operational evaluation" while the second type of evaluation can be classified as a "diagnostic evaluation". The modeled attainment test recommended in Section 3 uses models to predict the response of ozone to controls and then applies the resulting relative reduction factors to *observed* (rather than modeled) ozone. Thus, while historically, most of the effort has focused on the operational evaluation, the relative attainment test makes the diagnostic evaluation even more important.

In addition to the model performance evaluation, diagnostic analyses are potentially useful to better understand whether or not the predictions are plausible. Diagnostic analyses may also be able to provide: (1) information which helps prioritize efforts to improve and refine model inputs, (2) insight into which control strategies may be the most effective for meeting the ozone NAAQS, and (3) an indication of the "robustness" of a control strategy. That is, diagnostic tests may help determine whether the same conclusions would be reached regarding the adequacy of a strategy if alternative, plausible, assumptions were made in applying the model for the attainment test.

In this section, we first identify and discuss methods which may be useful for evaluating model performance. It is recommended that performance be assessed by considering a variety of methods. The section concludes by identifying several potentially useful diagnostic tests which States/Tribes should consider at various stages of the modeling analysis to increase the confidence in the model predictions of future ozone levels.

15.1 What are the Procedures for Evaluating An Air Quality Model?

As noted above, model performance can be assessed in one of two broad ways: how accurately does the model predict observed concentrations for specific cases, and how accurately does the model predict *responses* of predicted air quality to changes in inputs (e.g. relative reduction factors)? Given existing data bases, nearly all analyses have addressed the first type of performance evaluation. The underlying rationale is that if we are able to correctly characterize changes in concentrations accompanying a variety of meteorological conditions, this gives us some confidence that we can correctly characterize future concentrations under similar conditions. Typically, this type of operational evaluation is comprised principally of statistical assessments of model versus observed pairs. Operational evaluations are generally accompanied by graphical and other qualitative descriptions of the model's ability to replicate historical air quality patterns. The robustness of an operational evaluation is directly proportional to the amount and quality of the ambient data available for comparison. For the 8-hour ozone modeling, States/Tribes should compare all 1-hour observations and predictions (above a certain threshold), as well as all observed and predicted 8-hour daily maxima. Generally, if model performance is acceptable for the hourly pairs, one would expect the 8-hour performance to be acceptable as well.

The second type of model performance assessment, a diagnostic evaluation, can be made in several ways. One way to evaluate the response of the model is to examine predicted and observed ratios of "indicator species". If ratios of observed indicator species are very high or very low, they provide a sense of whether further ozone production at the monitored location is likely to be limited by availability of NOx or VOC (Sillman, 1995). Agreement between paired observed and predicted high (low) ratios suggests a model may correctly predict sensitivity of ozone at the monitored locations to emission control strategies. Thus, the use of indicator species has the potential to evaluate models in a way which is most closely related to how they will be used in attainment demonstrations. A second way for assessing a model's performance in predicting sensitivity of ozone to changes in emissions is to compare model projections after the fact with observed trends. Retrospective analyses provide potentially useful means for diagnosing why a strategy did or did not work as expected. They also provide an important opportunity to evaluate model performance in a way which is closely related to how models are used to support an attainment demonstration. More types of diagnostic analyses are provided in Section 15.3. We recommend that diagnostic analyses be performed during the initial phase of the model application and during any mid-course review.

15.2 How Should the Operational Evaluation of Performance Be Completed?

This section describes the recommended statistical measures and other analytical techniques which should be considered as part of an operational evaluation of ozone model performance. Note that model predictions from the ramp-up days should be excluded from the analysis of model performance. It is recommended that, at a minimum, the following three statistical measures be calculated for hourly ozone and 8-hourly maxima over the episode days in an attainment demonstration.

- Mean Normalized Bias (MNB): This performance statistic averages the model/observation residual, paired in time, normalized by observation, over all monitor times/locations. A value of zero would indicate that the model over predictions and model under predictions exactly cancel each other out. The calculation of this measure is shown in Equation 15.1.

- Mean Normalized Gross Error (MNGE): This performance statistic averages the absolute value of the model/observation residual, paired in time, normalized by observation, over all monitor times/locations. A value of zero would indicate that the model exactly matches the observed values at all points in space/time. The calculation of this measure is shown in Equation 15.2.

- Average Peak Prediction Bias and Error: These are measures of model performance that assesses only the ability of the model to predict daily peak 1-hour and 8-hour ozone. They are calculated essentially the same as the mean normalized bias and error (Equation 15.1 and 15.2), except that they only consider daily maxima data (predicted versus observed) at each monitoring location. In the attainment test, models are used to calculate relative reduction factors near monitoring sites by taking the ratio of the average 8-hour daily maximum concentrations calculated for the future and current cases. Thus, the model's ability to predict observed mean 8-hour daily maxima is an important indicator of model performance.

Equation 15.1

$$MNB = \frac{1}{N}\sum_{1}^{N}\left(\frac{(Model - Obs)}{Obs}\right)\cdot 100\%$$

Equation 15.2

$$MNGE = \frac{1}{N}\sum_{1}^{N}\left(\frac{|Model - Obs|}{Obs}\right)\cdot 100\%$$

EPA recommends that the three metrics above be calculated two ways: 1) for pairs in which the 1-hour or 8-hour observed concentrations are greater than 60 ppb[41], and 2) for all pairs (no threshold)[42]. This will help to focus the evaluation on the models ability to predict NAAQS-relevant ozone and minimize the effects of the normalization. In terms of pairing model predictions with monitored observations, EPA recommends that the grid cell value in which the monitor resides be used for the calculations. It would also be acceptable to consider bi-linear interpolation of model predictions to specific monitoring locations[43]. States/Tribes should

[41] Past ozone modeling applications have used a minimum cutoff of either 40 ppb or 60 ppb. Due to the interest in predicted ozone concentrations at or above the 8-hour standard (85 ppb), the higher cut off (60 ppb) is recommended.

[42] The use of a 0 ppb threshold can add valuable information about the ability of the model to simulate a wide range of conditions. Because of the tendency of the MNB and MNGE metrics to inflate the importance of biases at the lowest observed values (which are in the denominator), it is recommended that the alternate metrics of normalized mean bias (NMB) and normalized mean gross error (NMGE) be used as substitutes for evaluations with no minimum threshold.

[43] In certain instances, States/Tribes may also want to conduct performance evaluations using the “near the monitor” grid cell arrays. A “near the monitor” analysis may be useful when

recognize that, even in the case of perfect model performance, model-observed residuals are unlikely to result in exact matches due to differences between the model predictions which are volume averages and the observations which are point values.

The statistics should initially be calculated for individual days (averaged over all sites) and individual sites (averaged over all days). As appropriate, States/Tribes should then aggregate the raw statistical results into meaningful groups of subregions or subperiods.

Other statistics such as normalized mean bias, normalized mean gross error, fractional bias, fractional error, root mean square error, and correlation coefficients should also be calculated to the extent that they provide meaningful information (see Appendix A for definitions). Wherever possible, these types of performance measures should also be calculated for ozone precursors and related gas-phase oxidants (NOx, NOy, CO, HNO_3, H_2O_2, VOCs and VOC species, etc.) and ozone (and precursors) aloft.

Along with the statistical measures, EPA recommends that the following four sets of graphical displays be prepared and included as part of the performance analysis.

- Time series plots of model and predicted hourly ozone for each monitoring location in the nonattainment area, as well as key sites outside of the nonattainment area. These plots can indicate if there are particular times of day or days of the week when the model performs especially poorly.

- Scatter plots of predicted and observed ozone at each site within the nonattainment area (and/or an appropriate subregion). These plots should be completed using: a) all hours within the modeling period for hourly ozone, and b) all 8-hour daily maxima within the modeling period. It may also be useful to develop separate plots for individual time periods or key subregions. These plots are useful for indicating if there is a particular part of the distribution of observations that is poorly represented by the model[44].

- Daily tile plots of predicted ozone across the modeling domain with the actual observations as an overlay. Plots should be completed for both daily 1-hour maxima and daily 8-hour maxima. These plots can reveal locations where the model performs poorly. Superimposing

strong ozone gradients are observed, such as in the presence of a sea breeze or in strongly oxidant limited conditions. Furthermore, a “near the monitor” performance evaluation is consistant with the RRF methodology.

[44] Quantile-quantile (Q-Q) plots may also provide additional information with regards to the distribution of the observations vs. predictions. But due to the fact that Q-Q plots are not paired in time, they may not always provide useful information. Care should be taken in interpreting the results.

observed hourly or daily maximum concentrations on the predicted isopleths reveals useful information on the spatial alignment of predicted and observed plumes.

- Animations of predicted hourly ozone concentrations for all episode days or for certain periods of interest. Animations are useful for examining the timing and location of ozone formation. Animations may also reveal transport patterns (especially when looking at ozone aloft).

15.3 What Types of Analyses Can be Done to Evaluate the Accuracy of the Model Response: Diagnostic Evaluations?

This section lists possible analyses that could be performed to investigate the ability of the model to accurately forecast changes in ozone resulting from changes in ozone precursor emissions. States/Tribes are encouraged to complete as many of these types of analyses as possible, in order to increase confidence in the modeled attainment projections.

Observational models: In Section 5 it was noted that measurements of certain “indicator species ratios” are a potentially useful way to assess whether local ozone formation is VOC- or NOx-limited at any particular point in space and time. A performance evaluation which includes comparisons between modeled and observed ratios of indicator species (e.g., O_3/NOy, O_3/HNO_3) can help reveal whether the model is correctly predicting the sensitivity of ozone to VOC and/or NOx controls (Sillman, 1995 and 1998) and (Sillman, 1997 and 2002). If a model accurately predicts observed ratios of indicator species, then one can conclude with additional confidence that the predicted change in ozone may be accurate. One precaution with respect to the use of indicator species is that there may be a range of observed ratios for which the preferred direction of control is not clear. When this occurs, agreement between predictions and observations does not necessarily imply that the response to controls, as predicted by the model is correct. A second precaution is that application of this method often requires more measurements than are commonly made. Despite these precautions, comparing predicted and observed ratios of indicator species provides a means of assessing a model’s ability to accurately characterize the sensitivity of predicted ozone to changes in precursors.

Other observational methodologies exist and can be used in a similar manner. The Smog Production (SP) algorithm is another means by which ambient data can be used to assess areas that are NOx or VOC-limited (Blanchard et al., 1999). Additionally, it has been postulated that differences in weekend-weekday ozone patterns may also provide real-world information on which precursors are most responsible for ozone formation in any given area (Heuss et al., 2003). In areas where there are large differences between average weekend and weekday ambient ozone concentrations over the span of several seasons, it would be useful to compare statistical model performance for weekends versus weekdays. This would allow one to assess whether the model is capturing the effect of the emissions differences which are presumably driving the real-world concentration differences. This technique is not recommended if: 1) the number of days modeled is too few to result in an appropriate sample size of days, 2) there is no clear difference between ozone observations on the weekend versus weekdays, and/or 3) it is not possible to attribute differences in weekend/weekday differences to emissions differences. Despite these reservations associated with all of the various observational modeling approaches, these techniques allow one to evaluate the ability of the model to accurately predict changes in ozone concentrations. States/Tribes should include these comparisons in their efforts to evaluate model performance, whenever feasible.

Probing Tools: Recently, techniques have been developed to embed procedures within the code of an air quality model which enable users to assess the contributions of specific source categories or of specific geographic regions to predicted ozone at specified sites (Zhang et al., 2003). Various techniques have been implemented into various air quality models, but three of the most commonly used probing tools are photochemical source apportionment (Environ, 2004), the direct decoupled method (DDM) (Dunker, 1980 and 1981), (Environ, 2004) and process analysis (Jeffries, 1994 and 1997); (Jeffries, 1996); (Jang, 1995); (Lo, 1997). In the context of model performance evaluation, these attribution procedures are useful in that they allow one to “track” the importance of various emissions categories or phenomena contributing to predicted ozone at a given location. This can provide valuable insight into whether the model is adequately representing the conceptual description of ozone patterns in the nonattainment area. In the cases where model performance is subpar, these analyses can be useful for indicating where model input or model algorithm improvements are most needed.

Retrospective Analyses: A retrospective analysis is intended to examine the ability of the model to respond to emissions changes by comparing recent trends in observed ozone concentrations to the model-predicted trend over the same time period. As part of this analysis the model is run for current episodes and episodes in one or more historical time periods using the emissions and meteorological inputs appropriate for each time period modeled. While retrospective analyses may be useful, it may be difficult to obtain meteorological and emissions inputs for the historical time period(s) that are calculated using techniques and assumptions which are consistent with the calculation of these same inputs for the current time period. Using inconsistent inputs will confound the interpretation of the predicted trend. In Section 5, we noted that a retrospective analysis can be a useful tool for diagnosing why an areas has not attained the NAAQS. To that end, it is recommended that States/Tribes archive all modeling files and document assumptions and procedures used for calculating model inputs in order to facilitate replications of the modeled analyses at future dates.

Alternative Base Cases: In some cases it may be useful to evaluate how the response of the model to emissions reductions varies as a function of alternative model inputs or model algorithms. These types of tests can be used to assess the robustness of a control strategy. As an example, States/Tribes could consider the effects of assumed boundary conditions on predicted effectiveness of a control strategy. If the model response does not differ greatly over a variety of alternative plausible configurations, this increases confidence in the model results. The parameters for sensitivity tests can include, but are not limited to: different chemical mechanisms, finer or coarser grid resolution, meteorological inputs from alternative, credible meteorological model(s), different initial/boundary conditions, and multiple sets of reasonable emission projections. Sensitivity tests can and should be applied throughout the modeling process, not just when model performance is being evaluated. In cases where the operational model performance is considered to be poor, these tests may help indicate where base case input/algorithm changes are warranted.

15.4 How Should the Results of the Model Evaluation be Assessed?

In EPA guidance for the 1-hour ozone attainment demonstrations (U.S. EPA, 1991a), several statistical goals were identified for operational model performance. These goals were identified by assessing past modeling applications of ozone models and determining common ranges of bias, error, and accuracy (Tesche et al., 1990). The 1-hour guidance noted that because of differences in the quality of the applications considered, it was inappropriate to establish "rigid criterion for model acceptance or rejection" (i.e., no pass/fail test). It was recommended that these ranges should be used in conjunction with the additional qualitative procedures to assess overall model performance.[45]

With the additional experience of another decade of photochemical modeling, it is clear that there is no single definitive test for evaluating model performance. All of the tests identified in Sections 15.2 and 15.3 have strengths and weaknesses. Further, even within a single performance test, it is not appropriate to assign “bright line” criteria that distinguish between adequate and inadequate model performance. In this regard, EPA recommends that a “weight of evidence” approach (like that described in Section 4) be used to determine whether a particular modeling application is valid for assessing the future attainment status of an area. EPA recommends that States/Tribes undertake a variety of performance tests and weigh them qualitatively to assess model performance. Provided suitable data bases are available, greater weight should be given to those tests which assess the model capabilities most closely related to how the model is used in the modeled attainment test. Generally, additional confidence should be attributed to model base case applications in which a variety of the tests described above are applied and the results indicate that the model is performing well. From an operational standpoint, EPA recommends that States/Tribes compare their evaluation results against similar modeling exercises to ensure that the model performance approximates the quality of other applications.

[45] In practice, however, most 1-hour ozone modeling applications using the 1991 guidance tended to focus almost entirely on meeting the three statistical "goals" for bias, error, and accuracy at the expense of more diagnostic evaluation.